AWS Honeypot Attack Data Analysis

Kuldip Gadapa
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
kgadapa@gmu.edu

Satya Sai Jayanth Devineni
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
sdevine3@gmu.edu

Yeshwanth Reddy Bommu
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
ybommu@gmu.edu

Vineel Vishwanth Busi
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
vbusi@gmu.edu

***Abstract* - This study aims to analyze and predict the AWS Honeypot data, which comprises of various country source, host, prototype and time related data to know the exact attacking patterns of the hackers. Also, the paper draws some regression analysis about further attacks, that could be pre-assumed (known in advance). Several patterns like type of protocol used, locale, peak time of attack and top source address can be addressed specific to a country and target chosen. So, this study helps us to understand the critical information and apply higher protection to those sources that are most targeted. Using concepts of machine learning like regression and analysis techniques we retrieved several out of bound patterns and visualized to produce interesting outcomes from a huge observation. Our contribution can be analyzed about understanding threat intelligence and investigate major attacks in various countries. As a result, the outcomes can be used before preparing a secure infrastructure.**

Keywords - AWS Honeypot, regression analysis, machine learning, visuals, secure infrastructure, patterns.

I. INTRODUCTION

According to Online Digital industry experts, a Cyberattack is defined as a trial to obtain unauthorized and unaware access to one's personal and official assets (data, money) through network which is finally lead to misuse and destroy their assets. It is an offensive maneuver that effects computer information systems, computer networks, infrastructure, and personal computer devices. It can be implemented by nations, states, individuals, hackers, groups, or organizations. In our project, we focused on Amazon Web Services (AWS) honeypot attack data and visualizing those cyberattacks [1].

Honeypot aim at cyber criminals by gathering their information and targets. The tend to imitate a system to be true and rectify the existing vulnerabilities. The behavior of the hackers, malicious IP addresses, network protocols and other vital data are collected by honeypots. Honeypots are reliable, difficult to detect, and fast. They are platform independent, compatible and have low cost. They are helpful as they can route or intercept network traffic at any point of time.

Honeypot is an effective way of tackling the impact of cyberattack on computer infrastructure. It mimics a system to cause a cyberattack. It can be used to identify, deviate and know functionality of cyber criminals but it cannot stop attacks completely. They could be utilized as traps for cybercriminals as they think it is a legitimate target because it has system applications and data. AWS Honeypot resembles as Amazon network to outsiders which is maintained by Amazon IT teams. They monitor traffic for suspicious systems, track the attacks and operations to analyze what they want. Finally, they diagnose their security measures, firewall performances and perform necessary updates [1].

II. LITERATURE REVIEW

Amazon Web Service (AWS) honeypot is nothing but a trap point and security mechanism deliberated to tempt the attempted attack and if any source accesses the honeypot, the IP addresses will be recorded. Generally, a honeypot is the distraction for the attacker from their actual attack attempt and it will collect the information of the attacker by observing their request responses and the target hosts. Nowadays the cyber-attacks are immeasurable and more sophisticated to the companies, individuals, industries and government [2].

In 1986, the system admin of UC Berkeley named Clifford Stoll was involved in a process to track the charge for $0.75 of a Unix system at the lab. He used two honeypots to track the attacker. The actual target for the attacker was the nuclear defense secrets and later Clifford Stoll created a fictional department working on “Star Wars” to attract the attacker and he was later arrested. Since then honeypots became standard and the deception toolkit was launched in 1997. The honeynet project remained as the active security community resource [3].

There was a record of 451,581 attacks in a 6 months duration on AWS honeypots. AWS honeypot deals the attackers in a simple method by attracting the attackers with honeypot then the attacker will encounter the honeypot instead of our servers. The top 10 popular AWS data centers include Sydney, Sao Paulo, California, Mumbai, Frankfurt, London, Paris, Ireland, Singapore and Ohio were placed with the cloud server honeypots by an enterprise security company. Most of the honeypot projects are open source and there is a honeypot project that has extension tools where it will also analyze the data that will be collected by the honeypot [4].

Types of Honeypots:

The following are the different types of honeypots:
A low interaction honeypot is a Virtual Machine that just represents the frequently common attacks registered. The low interaction honeypot is very simple to create and maintain it, but the attacker will be able to know that it is not a genuine platform.

A high interaction honeypot will use virtual machines to keep the systems isolated. In this type, several honeypots can run in a single physical machine. For several honeypots it makes easier to scale up. By using high interaction honeypot, the researchers will be able to learn the tools that the attackers will use to attack a data which is private data and confidential.

A pure honeypot is a physical server which is configured in way to attract the attackers and there is a special software which monitors the connection between both the network and the honeypot platform. There are also some disadvantages with these type of honeypots as it requires intensive manpower to configure and maintain it [3].

III. PROBLEM DEFINITION

The key problem is to identify malicious activity that organizations tend to fortify. A honeypot is used for such purpose that will deliberately configure with known vulnerabilities at a location to make more tempting or obvious target for attackers. As honeypot has no production data or don’t participate in legitimate traffic on your network and that is how we can record and identify cybercrime. The definition covers a diverse array of systems, from simple virtual machines which offer a few vulnerable systems to build fake networks spanning multiple servers. The goals of honeypot are diverse as they can be used as defense in depth to academic research [3].

A honeypot is always isolated and monitored depending on the importance of the resources like attacker’s data, security mechanism, etc. It is like a trap that is set to observe, detect, mislead and divert an unauthorized user from his attempts on reaching an information system. Several companies must follow this bating procedure to safeguard the data from various threats and challenges like unidentified access during off time, using bad IP addresses for breaching, cross-site scripting attacks, distributed denial of service attacks and HTTP flood attacks [5].

Our main aim is to analyze and investigate the profiles of attackers, prone areas and the target. It helps to prepare a better defense by identifying important information and understand most attacks from a country and type of protocol used. Situations like cyber terrorist attacks on a nation data, research on data from honeypot on critical data, etc., We can also identify the domain to be protected prior to attack and rectify the method used by attackers.

Research honeypots allows close analysis of how hackers do their dirty work. The hacker’s techniques on using infiltrate systems, escalate privileges, etc. are scrutinized. They are set up by security companies, academics, and government agencies to examine the threat landscape. However, once the honeypot is detected its value diminishes and it is used by spamming industries to identify spam-catching honeypots [3].

IV. DATASET DESCRIPTION

Our dataset contains the attack data of the Amazon web services (AWS) containing the following data which include datetime, host, src, proto, type, spt, dpt, srcstr, cc, country, locale, locale abbr, postal code, latitude and longitude. Using this dataset, we can visualize the following which includes the geolocation of the attacked places, presenting the top attackers, detecting attacks by the host, and highly active IP addresses. It has only [6].

Nominal	Description
Datetime	Packet Arrival Date (YYYY-MM-DD)
host	Honeypot Server
src	Packet Source
proto	Packet Protocol Type
type	Packet Type
spt	Source Port
dpt	Destination Port
srcstr	Source IP Address
cc	Source Country Code
country	Source Country
locale	Source Location
localeabbr	Locale Abbreviation
postalcode	Postal Code

Ordinal	Description
latitude	Source Latitude
longitude	Source Longitude

This raw data will be then processed into a CSV file containing refined data about AWS honeypot. The dataset will have rows and 15 attributes.

V. METHODOLOGY:

From our dataset, we observed the data measurements (interval, nominal, ratio, ordinal) of our attributes. We would like to perform the quantitative research which is tested and objective. It also has both independent and dependent variables. For our hypothesis we want to use frequencies tables, cross tables and chi squared tests. We would like to use bar charts and line graphs to view the data in a simple way.

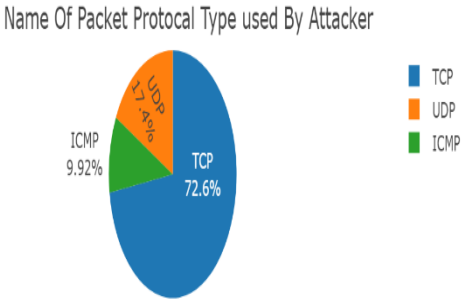
We had made our analysis from the Honeypot attach dataset to identify network attacks, security threats to understand and analyze some patterns, trends and characteristics. During our research, we fetched some related articles and dominant cyberattacks. We also observed reports that matched our attribute data and included some interesting observations.

A. Data Cleaning

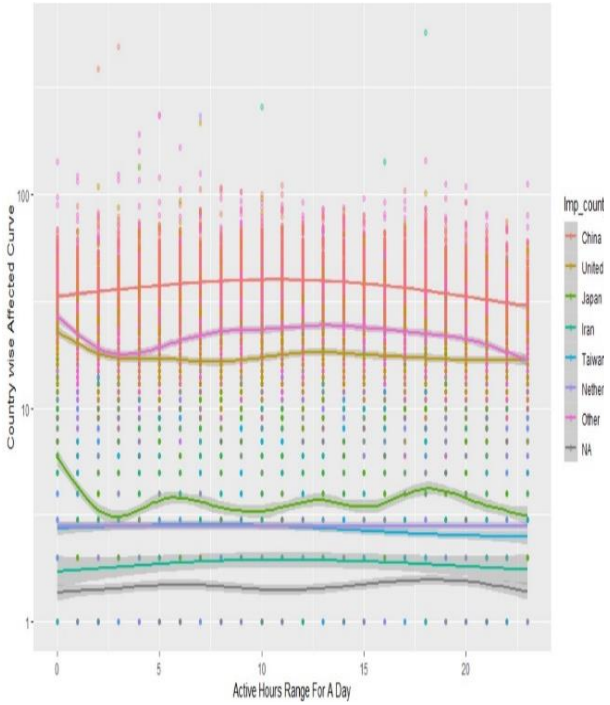
They were several missing values in the original data set that we used. We have filtered the data like geo-locations, the co-ordinates data was not full like postal, spt and dpt and we omitted some data that were not much important for analysis. As the data was huge it took a lot of time to re-examine the attributes and contents. The challenging task was he had over 9,55,000 data point with NA values in a 4,51,500 data row count and some data were wrongly recorded. We also converted the format of time on an hourly basis by splitting into 4 parts of the day.

B. Data Regression and Data Analysis

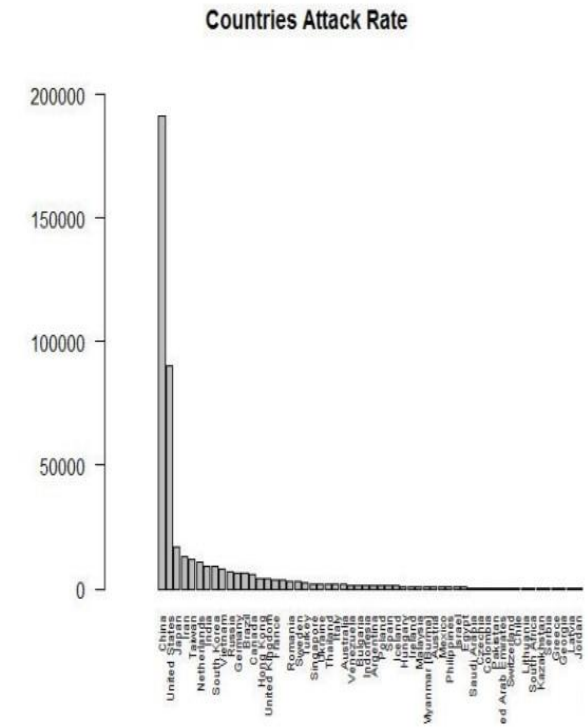
Firstly, we performed some quantitative co-relations on the common factors like type, src addresses, proto count, etc. We performed some generalizable conclusions. We also see some visuals like pie charts, interactive histograms, cross-tables, etc., and performed correlation analysis.



From the above plot we can observe the three different protocols TCP, UDP and ICMP used by the attacker. The most used protocol by the attacker is TCP and least used is ICMP.



The above plot shows that china has the most attack rate and is maintained through out the day. We can witness united states and Taiwan has most attacked cases in the morning and remains fluctuated.



As we observe the above plot, we can say that most of the false IP addresses and attacks are from China and united states.

C. Methods for correlation analyses

Pearson correlation (r), measures a linear dependence between two variables. It's also known as a parametric correlation test because it depends to the distribution of the data.
It can be used only when x and y from a normal distribution. The plot of y = f(x) is linear regression curve. Kendall tau and Spearman rho are rank-based correlation coefficients (non-parametric). Pearson correlation is the most practiced method [7].
Correlation Formula: In the formula below,

x and **y** are two vectors of length **n**
mxmx and mymy corresponds to the means of x and y, respectively.

Pearson Correlation Formula
$r=\sum(x-mx)(y-my)/\sum(x-mx)2\sum(y-my)2$
mxmx and mymy are the means of x and y variables.
The p-value (significance level) of the correlation can be determined:

1. by using the correlation coefficient table for the degrees of freedom: df=n−2, where n is the number of observations in x and y variables.
2. or by calculating the **t value** as follow:

$$t=r^*\sqrt{n-2}/\sqrt{1-r2}$$

In the case 2) the corresponding p-value is determined using **t distribution table** for df=n−2. It the p-value is <5%, then the correlation between x and y is significant [7].

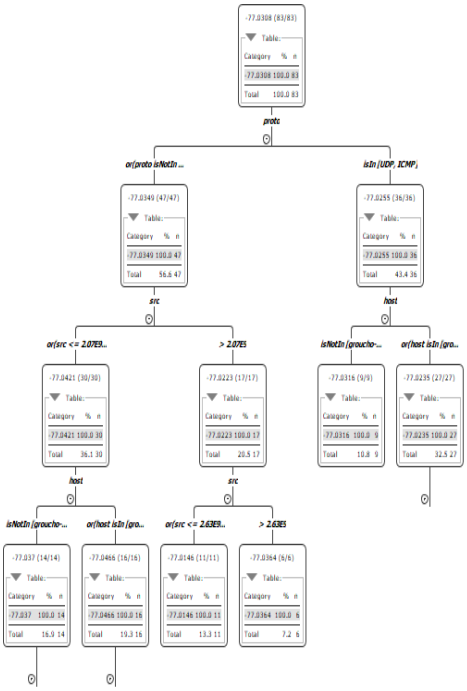
```
> chisq.test(HP$host,HP$proto)

        Pearson's Chi-squared test

data:  HP$host and HP$proto
X-squared = 36246, df = 16, p-value < 2.2e-16

> chisq.test(table(HP$country))

        Chi-squared test for given probabilities

data:  table(HP$country)
X-squared = 17678993, df = 177, p-value < 2.2e-16

> chisq.test(HP$country,HP$proto)

        Pearson's Chi-squared test

data:  HP$country and HP$proto
X-squared = 165161, df = 354, p-value < 2.2e-16
```

According to the chi square results, the p value is less than 0.05 and with 95% confidence, we can say: There is a relationship between time of attack, protocol type, target host and attacker country.

sn$host	sn$prot TCP	UDP	ICMP	Row Total
groucho-eu	17405	4380	2169	23954
	0.003	9.686	18.199	
	0.039	0.010	0.005	
groucho-norcal	16421	4823	3322	24566
	113.281	67.394	320.773	
	0.036	0.011	0.007	
groucho-oregon	84179	7755	2142	94076
	3676.656	4566.163	5542.776	
	0.186	0.017	0.005	
groucho-sa	17112	4429	2775	24316
	17.074	8.247	54.337	
	0.038	0.010	0.006	
groucho-singapore	61024	6165	10962	78151
	319.951	4091.331	1326.191	
	0.135	0.014	0.024	
groucho-sydney	17177	4479	2800	24456
	19.320	10.595	57.391	
	0.038	0.010	0.006	
groucho-tokyo	72809	37285	16095	126189
	3874.440	10593.667	1019.573	
	0.161	0.083	0.036	
groucho-us-east	24663	4643	2473	31779
	108.343	146.398	146.836	
	0.055	0.010	0.005	
zeppo-norcal	17201	4820	2073	24094
	5.105	90.501	42.264	
	0.038	0.011	0.005	
Column Total	327991	78779	44811	451581

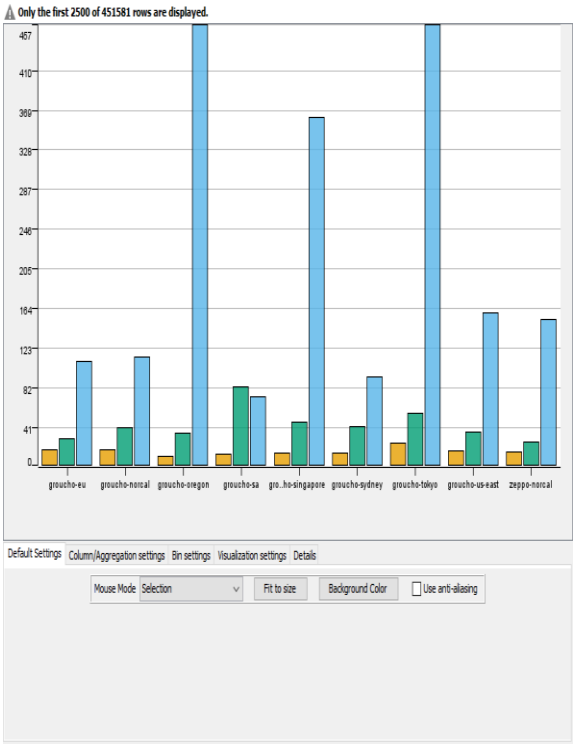
D. Methods for Regression and Clustering

1. Random Forest Regression- Learns a random forest* (an ensemble of decision trees) for regression. Each of the regression tree models is learned on a different set of rows (records) and/or a different set of columns (describing attributes), whereby the latter can also be a bit/byte/double vector descriptor (e.g. molecular fingerprint). The output model describes an ensemble of regression tree models and is applied in the corresponding predictor node using a simple mean of the individual predictions. Displays an interactive histogram view with different viewing options. The interactive histogram supports hilting and the changing of the x axis and aggregation column on the fly.

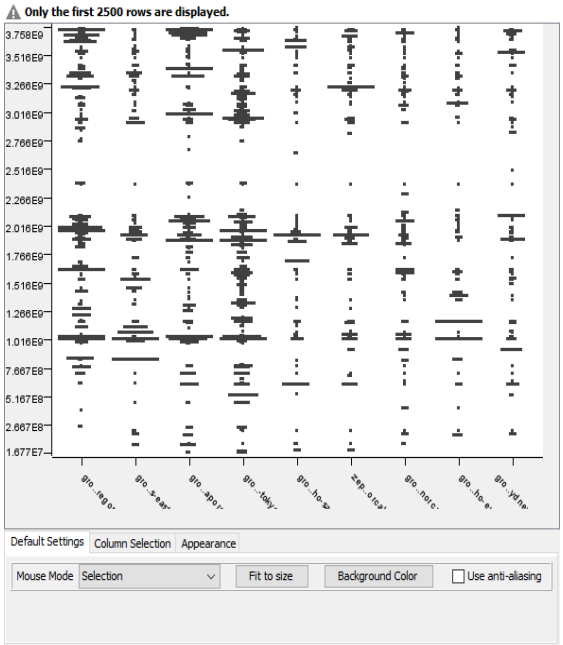


The above decision tree is the outcome of the applied regression using random forest method. It characterizes the dependency starting from the type of protocol used and source addresses. We can finally see that most of the country and host locations depend on the source of the attacker.

The below graph is an interactive histogram that is produced from the random forest sample. This is performed using KNIME software.



2. FUZZY C-MEANS- The fuzzy c-means algorithm is a well-known unsupervised learning technique that can be used to reveal the underlying structure of the data. Fuzzy clustering allows each data point to belong to several clusters, with a degree of membership to each one. Make sure that the input data is normalized to obtain better clustering results. The list of attributes to use can be set in the second tab of the dialog. The first output data table provides the original data table with the cluster memberships to each cluster. The second data table provides the values of the cluster prototypes. Additionally, it is possible to induce a noise cluster, to detect noise in the dataset, based on the approach from R. N. Dave: 'Characterization and detection of noise in clustering'. Creates a scatterplot of two selectable attributes. Then each datapoint is displayed as a dot at its corresponding place, dependent on its values of the selected attributes. The dots are displayed in the color defined by the Color Manager, the size defined by the Size Manager, and the shape defined by the Shape Manager.



KNIME, the Konstanz Information Miner, is a free, has integration platform for reporting and open source data analytics tool. The concept of modular data pipelining is adapted via ML and data mining components. It uses Nodes as GUI that blend several data sources, preprocessing (Extraction, Transforming and Loading) for mining, modelling, analysis and visualization with minimal programming. KNIME, open for innovation can be easily understood and downloaded to build workflows, projects and develop insights [8].

VI. CONCLUSION

We conclude that in this project we have tried to deduce the data of the honeypot. We applied some methodologies and used chi squared test. Also, random forest and fuzz-c-means are performed using KNIME. The obtained information from the results gives valuable information which includes the following as we have observed that there are some protocols which are used effectively in some systems and these systems have to be configured accordingly and maintained. Depending on the time of day and protocol the attacking countries vary with these factors, so we have to monitor the IP addresses from those countries. Some prevention measures should be taken for the protocols which are being attacked on specific days and months. These are the essential measures observed from our results.

References

[1] "AWS Honeypot Data: Visualizing The Threat of Cyberattacks," 2020. [Online]. Available: https://www.sisense.com/whitepapers/gofigure-aws-honeypot-data-visualizing-the-threat-of-cyberattacks/.

[2] "Cybersecurity," 2018. [Online]. Available: https://wiki.smu.edu.sg/1718t3isss608/Group01_Report.

[3] "What is a honeypot? A trap for catching hackers in the act," 2019. [Online]. Available: https://www.csoonline.com/article/3384702/what-is-a-honeypot-a-trap-for-catching-hackers-in-the-act.html.

[4] "Cybercriminals attack cloud server honeypot in 52 seconds," 2019. [Online]. Available: https://www.cio.com/article/3515424/cybercriminals-attack-cloud-server-honeypot-in-52-seconds.html.

[5] "eVitamins Case Study," 2017. [Online]. Available: https://aws.amazon.com/solutions/case-studies/evitamins/.

[6] "AWS Honeypot Attack Data," 2018. [Online]. Available: https://www.kaggle.com/casimian2000/aws-honeypot-attack-data.

[7] "Correlation Test Between Two Variables in R," 2020. [Online]. Available: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r.

[8] "End to End DATA sCIENCE," 2020. [Online]. Available: https://www.knime.com/.